Scalability

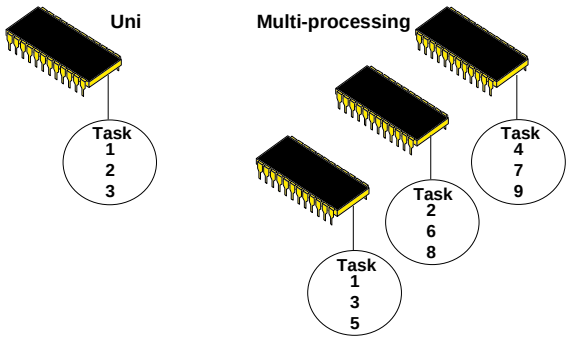
What is Scalability?

Computer customers of the past often bought mainframes twice the size they needed in anticipation of growth. They knew that if their business continued to grow they would grow into the machine - eventually. Customers today can buy computers to fit the size they need at the moment, and add more equipment as their business needs demand. They can take advantage of scalability in today's systems, driven in large part by the rapid advances in software and microprocessor technology.

A machine that is scalable has the ability to grow in size and speed. Some machines offer limited scalability by design, while some can grow to virtually any size needed. One approach to computing is uni-processing, in which one CPU performs all the application processing. An example of uni-processing is the typical PC. But many of the demanding applications being developed today require more power than any single processor can offer.

A computer that utilizes more than one processor is called a multiprocessor computer. Multiprocessing machines usually utilize different processors to perform different tasks as shown in the figure below.

Uni-processing vs. Multiprocessing

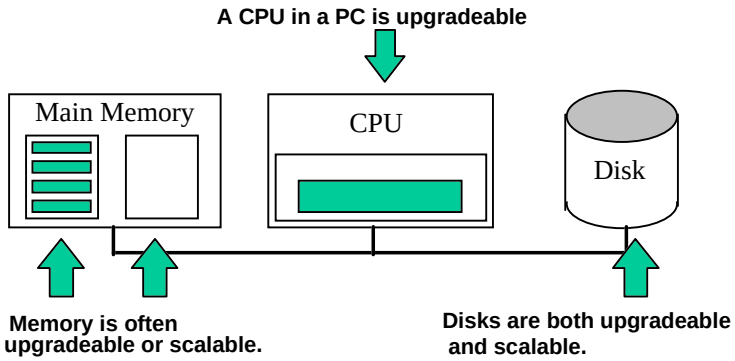


Upgrading vs. Scaling

What is the difference between an upgrade and scaling? An upgrade is the ability to replace a component with a faster or better component, whereas a machine that is scalable allows the user to add to or build on that component. For example, a PC normally has only one processor that you might be able to upgrade with a simple replacement. You do not have the ability to scale your processor because most PCs are not designed for two or more CPUs.

Memory is one of the areas in which PCs have the ability to scale. In order to increase memory capacity or speed you would normally want to add to your existing memory as opposed to replacing it. In this case, PCs are designed to be scaled with their additional slots for handling increased memory. The same holds true for your disk space. You could upgrade to a larger or faster disk, or you could scale by adding additional hard drives.

The figure below shows the components of a typical PC. Some of the components can be upgraded, some can scale, and some have the ability to do both.

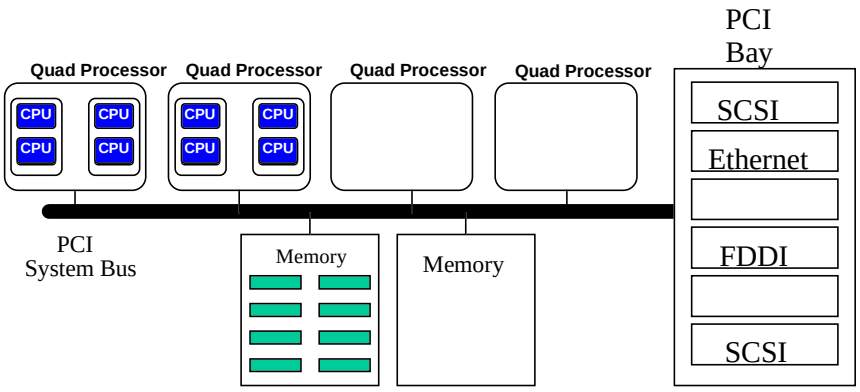


Symmetrical Multiprocessing

A Symmetric Multiprocessing (SMP) machine is a computer that utilizes multiple processors. These processors share memory and one copy of the operating system. SMP machines can scale by starting small with only two processors and adding more processors as business needs and applications grow. These computers have the ability to scale memory, cache, or disks as well as CPUs. Currently SMP machines are designed to scale from two to 32 processors.

There are limiting factors when dealing with SMP. You might think that you should be able to scale far above 32 processors. If you were to start with two processors and add two more, a near 100% improvement may result, but because there is only one operating system and all memory is shared, a progressive diminishing return will be realized as more processors are added. Most SMP systems will show worthwhile improvements until they scale above eight processors. This scaling formula also changes based on the operating system and the application. While Unix systems with sixteen or more processors are not uncommon today, Windows NT scalability is commonly thought to be limited to about four CPUs. Additionally, many operating systems or database applications can only utilize the first two gigabytes of memory. No one wants to pay for memory if it is not being used. As operating systems and databases improve their ability to scale, customers will be able to add the additional horsepower to take advantage of these improvements.

The picture below shows an SMP machine. Notice the ability to double the number of processors, double the memory, and add SCSI disk or network cards.



Linear Scalability through Parallel Processing

Some of the largest and most scalable systems in the world utilize parallel processing technology. Parallel processing takes SMP a step further by combining multiple SMP “nodes” that can work in parallel on a single application, usually based on a database that is fully “parallel-capable.” Because each node has its own copy of the operating system, and the nodes communicate through a specialized interconnect, adding additional nodes does not increasingly tax a single OS. Therefore, parallel processing can scale to much higher numbers in a more linear fashion than SMP alone.

An example of a very large parallel processing system is a well-known large retailer uses a system with over 700 processors on a database of up to 24 terabytes of online storage! The figure below shows a logical view of how one database table is spread among multiple processors.

Logical view of one SMP node using parallel processing

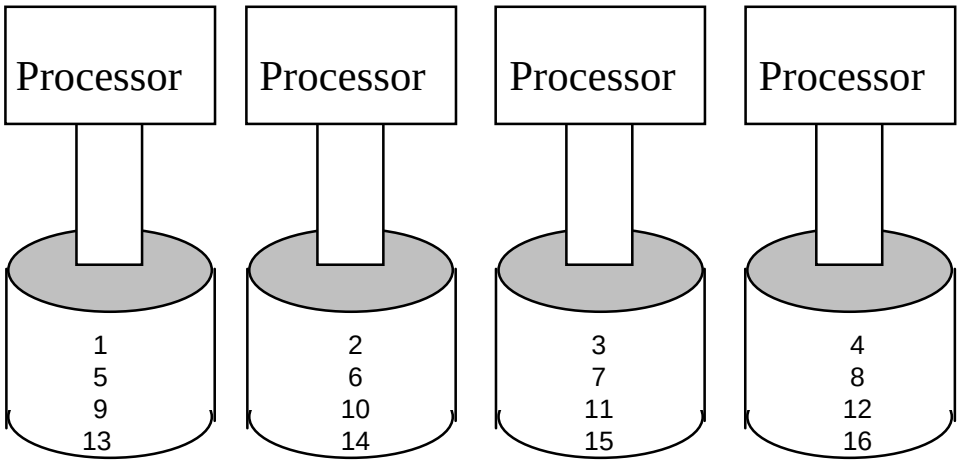


Table rows are spread evenly among processors and queries are processed in parallel

Key Points to Remember:

- Scalability is the ability to add additional processors, memory, or disks to increase performance, strength, or database size.
- SMP machines can scale processors, memory, and disks, but are limited in degree to which they may scale.
- Limitations are based on physical expansion capabilities and capacity of the operating system or database to take advantage of additional processors or memory.
- Parallel processing allows for more linear scalability than SMP systems.